Hindawi
Scientific Programming
Volume 2021, Article ID 3058141, 8 pages
https://doi.org/10.1155/2021/3058141

Research Article

Aerobics Action Recognition Algorithm Based on Three-Dimensional Convolutional Neural Network and Multilabel Classification

Qian Wang[1] **and Mingzhe Wang**[2]

[1]*College of Art, Xi'an Physical Education University, Xi'an 710068, China*
[2]*Mechanical and Electrical Department, Hebei Vocational College of Rail Transportation, Shijiazhuang 050000, Hebei, China*

Correspondence should be addressed to Mingzhe Wang; 107057@tea.xaipe.edu.cn

Received 20 April 2021; Accepted 16 June 2021; Published 5 July 2021

Academic Editor: Shah Nazir

In the context of modern people increasingly paying attention to health and promoting aerobics, the amount of data and audiences of aerobics videos has grown rapidly, and its potential application value has attracted widespread attention from scientific research and industry perspectives. This article has integrated computer vision and deep learning related knowledge to realize the intelligent recognition and representation of specific human movements in aerobics video sequences. The study proposes an automatic recognition method for floor exercise videos based on three-dimensional convolutional networks and multilabel classification. Since two-dimensional convolutional neural networks (CNNs) lose time information when extracting features, so to overcome this, the proposed research uses three-dimensional convolutional networks to perform video recognition. The feature is taken in time and space, and the extracted features are subjected to multiple binary classifications to achieve the goal of multilabel classification. Various comparison and simulation experiments are conducted for the proposed research, and the experimental results prove the effectiveness and superiority of the approach.

1. Introduction

With the rapid development of related technologies such as computers [1–3], networks [4–6], and multimedia, multimedia data have shown an exponential growth trend. Video [7] is a common form of multimedia data, and it is also an important part of multimedia data, which is closely related to our daily lives. The video contains the most abundant data information, with a complex structure and a large amount of data. Faced with such a huge amount of video data, automatic video description can better manage and utilize these rich video resources and can help users improve the indexing speed and search quality of online videos so that they can play a greater role. For visually impaired people, through the automatic description of the video combined with text-to-speech technology, the text in the computer is converted into continuous natural language for communication [8]. It can help them better understand the content in the video, thus making the life of the visually impaired more convenient. In the field of video automatic description research, video automatic analysis and understanding based on human actions has gradually become a hot research problem in the field of computer vision and pattern recognition in recent years. It has a wide range of application prospects in the fields of intelligent life assistance, advanced human-computer interaction, and content-based video retrieval [9] and is closely watched by researchers at home and abroad.

Faced with the low-level video features in the current aerobics video analysis research that cannot accurately reflect human high-level semantic concepts, the action recognition algorithm in traditional RGB video has high time complexity and low recognition accuracy, and the use of a single feature cannot meet the massive amount of existing

video data. Growth of complex and other issues, its automatic description research has important theoretical research significance and extensive practical application value. In terms of theoretical research, the research on automatic video description of floor exercise is a cross-cutting subject that integrates machine learning, pattern recognition [10–12], video analysis, computer vision, and cognitive science and provides a good basis for research in these fields. In-depth research can promote the development of related disciplines.

Regarding the problem of automatic identification of aerobics videos [13], which is a difficult point in visual research, in practical applications, the research of automatic identification based on aerobics videos has a wide range of application prospects and potential economic value. In addition to the abovementioned video retrieval and convenience to the visually impaired, potential application areas include sports assisted training, human-computer interaction, and project promotion. First of all, it can satisfy human-computer interaction [14]. In the complex floor exercise set of movements, it is particularly important to identify various human movements quickly. When watching aerobics competitions, commentators often have problems such as delay and error in interpretation of decomposed movements. In this paper, we strive to achieve a higher recognition accuracy in the automatic understanding of human movements based on video and even realize real-time movement recognition and interpretation. For non-professionals, if automatic recognition can be realized, it can not only improve the feeling of watching the game but also make it more convenient for them to understand and learn aerobics.

Secondly, it can assist sports training. In the aerobics video, the movement of the human body is very complex with strong skills. Compared with the daily exercise, the analysis of aerobics video is more difficult and challenging. The analysis of aerobics videos can not only bring more watching effects to sports games but also help coaches to analyze the games and assist athletes in training. Through the research on the automatic understanding of aerobics, while improving the accuracy of aerobics movement recognition, this paper analyzes the movement data so as to excavate the regularity characteristics of gymnastics technology innovation and development and realize the function of auxiliary training. For example, taking related athletes as the main research object, the paper analyzes the differences between the difficulty, arrangement, and quality of the complete sets of movements between the winners and ordinary athletes, studies the development and innovation trend of aerobics, and adjusts the training countermeasures so as to improve the skill level of athletes [15].

Finally, project promotion can be carried out. Taking aerobics as a typical research object, knowledge transfer can be used to effectively identify and locate human movements in aerobics videos. By referring to the method of aerobics movement recognition, it can be applied to other sports so as to expand the research results. Following are the main innovations points of this paper:

(i) To improve the existing algorithm model, the accuracy of automatic aerobics recognition is improved.

(ii) Automatic aerobics video recognition is transformed into a multilabel classification problem. In order to extract the temporal and spatial feature representation in the video, a three-dimensional CNN [16–18] is used as a feature extractor. Then, a two-class classifier for a single decomposition action is constructed, and each video will perform two-class calculations for all categories to complete the multilabel classification process.

(iii) To conduct comparison and ablation experiments, the experimental results prove the effectiveness and superiority of our algorithm.

2. Background

2.1. Convolutional Neural Network. A typical CNN schematic diagram is shown in Figure 1. It consists of three parts: the first part is the input layer, the second part is the several hidden layers, and the third part is the one output layer. Each layer is composed of multiple neural units. CNN [19, 20] has two key ideas, which determine its performance in solving problems related to computer vision field which is particularly outstanding. The first point is that CNN makes use of the two-dimensional structure of images. Since pixels in adjacent areas are usually highly correlated, CNN does not need to establish one-to-one connection between pixel units like traditional neural networks but can directly use grouped local connections. The second point is that the CNN architecture relies on feature sharing, where each channel is generated by convolution using the same filter at all locations.

In the specific CNN network structure, the hidden layer usually includes a convolutional layer, an activation function, a pooling layer, and a fully connected layer. The function of the convolutional layer is to extract the features of the input layer. It is composed of many convolutional units, and the parameters of the convolutional unit are optimized through the backpropagation of the convolutional network. In the process of recognition, the human brain first perceives each feature locally and then comprehensively sorts the local features to obtain global information. Therefore, the feature extraction of the convolutional layer plays a central role in the CNN. A CNN usually contains multiple convolutional layers. The shallow convolutional layer usually can only extract lower-level features. Commonly used CNN usually uses multiple layers in order to obtain deeper feature maps. Convolutional layer is used to iterate. The function of the activation function is to increase the nonlinear segmentation ability of the network. As an activation function, it generally satisfies the properties of nonlinearity, continuous differentiability, monotonicity, best unsaturated range, and approximate linearity at the origin. Commonly used activation functions include ReLU and Maxout. The pooling layer is also called the downsampling layer, usually after the convolutional layer. The

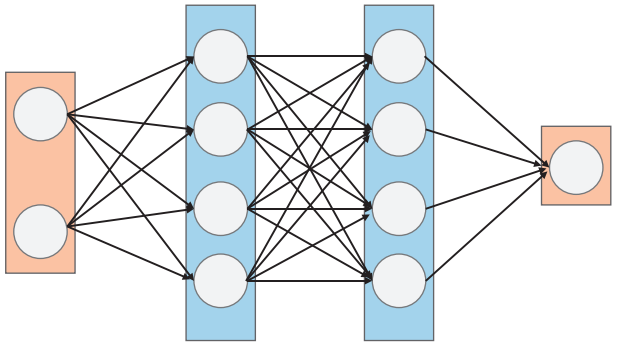

Figure 1: Schematic diagram of a typical neural network.

feature dimensions obtained by the convolutional layer are relatively large, and compressing and sampling the feature maps obtained by the convolutional layer can not only reduce the computational complexity of the network and improve the recognition of features but also avoid the overfitting problem to a certain extent. Common pooling methods include average pooling and maximum pooling. The fully connected layer will connect all the features by weighting, and the output value obtained is used in the calculation of the classifier.

2.2. Recurrent Neural Network. Recurrent neural network (RNN) [21–24] is a special neural network structure inspired by human beings' reliance on past experience and memory in the cognitive process. RNN is called a recurrent neural network. RNN not only gives the input of the previous moment the memory function but also gives the input of the next moment referring to the memory of the previous moment; that is, the current output of a sequence is composed of its input and the previous sequence. The output is jointly determined. The specific process performance will be applied to the previously memorized output when calculating the current output. RNN is different from CNN. In RNN, the input data have a time sequence, thus forming a sequence. This is the most critical point that distinguishes RNN from other neural networks [25–27], and it is also the fundamental reason why the "loop" can be established. The nodes between the hidden layers are unconnected in CNN and become connected in RNN, and the input of the hidden layer includes the output of the input layer and the output of the hidden layer at the previous moment. The hidden layer of the simplest structure of RNN is expanded in time, and its structure is shown in Figure 2. X represents the input sample, 0 represents the output, U and U, respectively, represent the weight of the sample input and output at the moment, t represents the time series, and the memory of the input sample at time t is expressed as follows:

$$S_t = f\left(W * S_{t-1} + U * X_t\right), \tag{1}$$

where W represents the weight entered at the previous moment.

In the actual application process, with the deepening of the network model, the problems of gradient explosion and gradient disappearance appear when the RNN model is

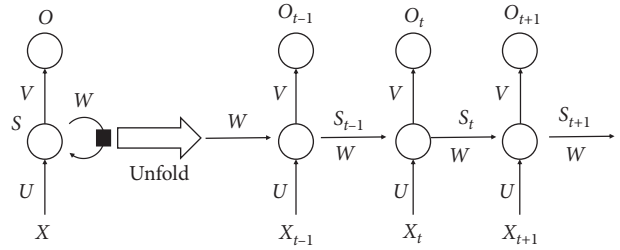

Figure 2: Hierarchical expansion diagram of RNN hidden layer.

trained often trouble researchers. Once the gradient disappears or the gradient explodes, the transfer performance of the training gradient will be greatly reduced, and the original purpose of the RNN model design cannot be achieved. That is, the training gradient cannot be transmitted in a long sequence, which eventually leads to a large deviation in the detection accuracy of the long sequence by the RNN. In order to achieve the long-term dependence problem that needs to be achieved during the training of the RNN model, a long- and short-term memory network is proposed. This network model improves the traditional RNN model by introducing a memory unit and a gate control memory unit. The memory unit can store historical information and the network. In the long-term state, the gate control determines the flow of information through linear intervention, which can selectively increase or decrease the transmission of information. The hidden layer of LSTM [28–31] is different from RNN, and its internal structure is more complicated.

3. Methodology

3.1. Attention Mechanism. The human brain pays attention to different parts of the brain differently when processing signals, known as the visual attention mechanism [32, 33]. Human vision can quickly scan the global image to obtain the target area that needs to be focused on, which is generally known as the focus of attention and then invest more attention resources in this area to obtain more detailed information of the target that needs to be focused on and suppress other useless information. The reason why this paper needs to use the attention mechanism is very intuitive. The decisive video frame for automatic description of the decomposing movements of floor exercises should be the method, direction, and angle of the athlete's body turning, and the weight of these video frames should be greater. This paper uses an attention mechanism which allows the decoder to weight each time feature vector of floor exercise video. Figure 3 shows the network structure after the attention mechanism is introduced.

This paper adopts the dynamic weighted sum of time feature vectors, and the formula is as follows:

$$\phi_t(X) = \sum_{i=1}^{n} \alpha_i^{(t)} x_i, \tag{2}$$

where $\sum_{i=1}^{n} \alpha_i^{(t)} = 1$, $\alpha_i^{(t)}$ is the proportion of the matching score between the hidden layer output and the entire video

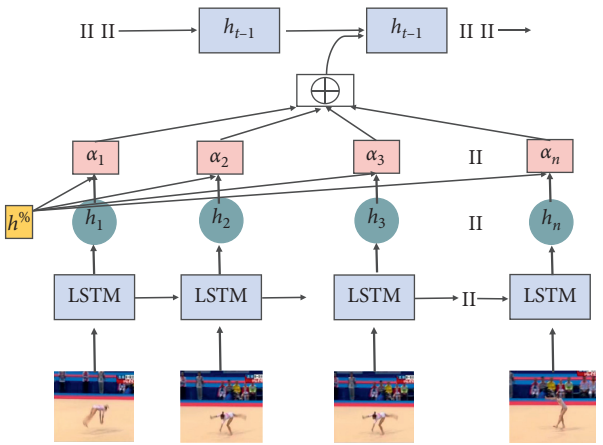

Figure 3: Diagram of attention mechanism.

representation vector at the moment in the overall score, and the calculation formula is as follows:

$$\alpha_i^{(t)} = \frac{\exp\left(\text{score}\left(x_i, h_i\right)\right)}{\sum_{j=1}^{n} \exp\left(\text{score}\left(x_i, h_j\right)\right)}, \tag{3}$$

where $\text{score}(x_i, h_i)$ represents the score value of the output h_i of the i-th hidden layer in the video feature vector x_i. The larger the score, the greater the attention of the input at this moment in the video, and its calculated as follows:

$$\text{score}\left(x_i, h_i\right) = \omega^T \tanh\left(W x_i + U h_i + b\right), \tag{4}$$

where ω, W, and U are the weight vectors and b is the bias.

3.2. Network Framework. In this chapter, the automatic recognition problem of aerobics videos is transformed into a video multilabel classification problem, and finally the classification results are further transformed into real floor exercise recognition. In order to achieve this process, the basic framework used in this section is shown in Figure 4. The framework in the figure can be divided into two parts based on the three-dimensional convolutional network to extract the multilabel video features of aerobics, and then SVM is used to extract multilabel classification that is performed on the pictures, and finally the mapping of the results of multilabel classification to natural language is completed, and the automatic description of the aerobics video is finally completed.

3.2.1. Feature Extraction. Compared with 2D convolutional networks, 3D convolutional networks can better model time information through 3D convolution and 3D pooling operations. In a two-dimensional convolutional network, the process of convolution and pooling is completed in space. In a three-dimensional convolutional network, they perform in time and space. In the introduction of 3D convolutional network above, it was proposed that images should be output when 2D convolutional network is processing images, and images should also be output when multiple images (which are regarded as different channels) are operated. Therefore, the time information of input data will be lost after each convolution operation in the two-dimensional convolutional network. Only three-dimensional convolution can preserve the time information of the input signal and produce the output quantity. The same principle can be applied to 2D pooling and 3D pooling.

3.2.2. Multiclassification of Video Based on SVM. After the video features are obtained, this article will establish a two-class classifier for each decomposition action to determine whether the video contains this type of action. For the establishment of the second-class classifier, the SVM classifier is used in the work of this article. In order to obtain the optimal linear interface of the SVM classifier, the basic idea of solving the problem is to transform the input space into a high-dimensional feature space through nonlinear transformation, which can be regarded as a linear classifier in a broad sense. As shown in Figure 5, the two types of training samples in the figure are represented by "∗" and "⌖" respectively, x_1 and x_2 represent the two feature items of the sample, and H is the interface of H', H_1, and H_2, respectively, which represent the closest to the interface of the two types of samples. The point is parallel to the plane of the interface. In order to ensure that the empirical risk is minimized in the support vector classification model, not only the optimal dividing line is required to correctly separate the two types of data but also the two types of classification interval (M in the figure) must be maximized. Therefore, although H' is also a boundary that can be classified correctly, it is not suitable as a boundary. The principle of interface selection is to make the support vector machine show better generalization ability.

Use $\{(a_1, c_1), \ldots, (a_N, c_N)\}$ to represent the linearly separable sample set of the two types of problems, where $a_i \in R^d$ and d represents the dimension. The category label $c_i \in \{-1, 1\}, i \in [1, N]$, w is a d-dimensional vector, and b is a constant. From this, the linear discriminant function can be obtained as follows:

$$c(a) = w^T a + b. \tag{5}$$

In order to obtain the maximum classification interval M, the interface needs to meet the following requirements:

$$w^T x + b \begin{cases} > \dfrac{M}{2}, & \text{for } y_i = 1, \\ < \dfrac{M}{2}, & \text{for } y_i = -1. \end{cases} \tag{6}$$

Normalize formula (6) so that all samples can satisfy $|c(a)| \geq 1$, and the sample with the smallest distance from the interface satisfies $|c(a)| = 1$; thus,

$$c_i\left(w^T a_i + b\right) \geq 1. \tag{7}$$

It can be deduced that $M = 2/\|W\|$. When $\|W\|$ is the smallest, the classification interval is the largest. And in order to satisfy that the objective function becomes a

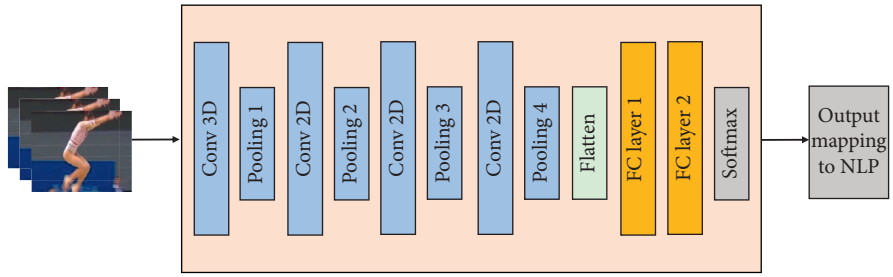

Figure 4: Network framework.

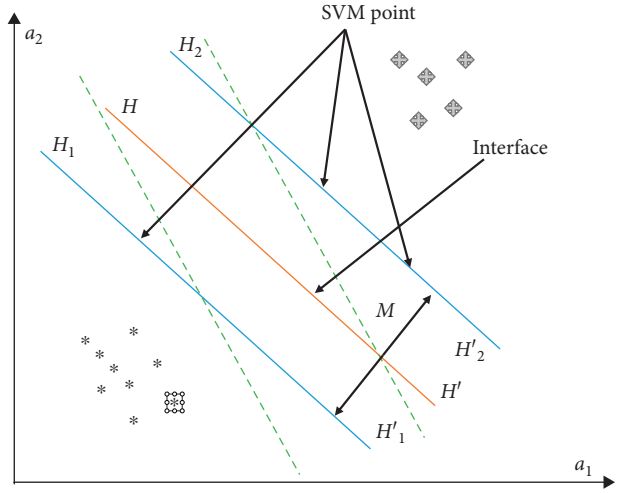

Figure 5: Schematic diagram of SVM interface.

quadratic programming problem, the minimum value of $\|W\|^2$ is taken. Thus, we can get

$$\min_{w,b} \frac{1}{2} w^T w$$
$$\text{s.t.} c_i\left(w^T a_i + b\right) \geq 1. \tag{8}$$

According to the Lagrange method, the corresponding Lagrange function is obtained:

$$L(w, b, a) = \frac{1}{2} w^T w - \sum_{i=1}^{N} a_i\left(c_i\left(w^T a_i + b\right) - 1\right). \tag{9}$$

The previous description is the case of linear separability, and the actual problems that need to be dealt with are often linear inseparable data sets, that is, nonlinear separable problems. Therefore, some misclassifications will inevitably occur in the classification process. One of the solutions is to transform the nonlinearity into a linearly separable problem. Specifically, the kernel function is introduced to map the nonlinearly separable problem in the input space to a higher-dimensional feature space. The Gauss radial basis kernel function is used in this chapter:

$$K(a_1 a_2) = \exp\left(-q\left\|a_1 - a_2\right\|^2\right). \tag{10}$$

For N two-classification problems of N decomposition actions, N two-class SVMs are constructed, and each SVM is trained in a one-to-many manner; that is, the video containing the current decomposition action is taken as a positive sample, and other data are taken as a negative sample.

3.2.3. Automatic Recognition Based on Multiple Classifications. In order to form a contrast experiment, this chapter maps the multilabel classification results into automatic recognition statements of aerobics. Different from the previous operation, this step does not involve the processing of video data but only compares the classification results with the test descriptions marked in Chapter 3. Through multiple binary classification SVM classifier, each video will get multiple 0-1 classification results, to identify the category of the assembled into a sentence because most of the video data are only less categories, and the logical relationship between each category is not obvious, so here temporarily do not consider the semantic information, The classification results are directly connected to form the description statement of aerobics movements, and the process of automatic recognition is completed.

4. Experiments and Results

4.1. Experimental Setup. The experiment in this section is carried out on the operating system application Ubuntu 16.04 version. The code to realize the experiment is based on TensorflowL.6.0 framework, and the language used is Python 2.7. The network model was trained on two Nvidia Titan 1080 graphics cards with 11 GB of memory. The input video data are sampled every 5 frames. The input of the C3D feature extraction model is a 16-frame long segment with 8 frames overlapping between two continuous segments. The FC1 activation of these segments is averaged to obtain a 4096-dimensional video descriptor.

4.2. Aerobics Multicategory Data Set. This paper collects a large number of high-standard events of professional athletes, including the Olympic Games, the World Championships, the National Games, and other male and female heavyweight events. First of all, these games are preprocessed. A complete aerobics video is completed by the participation of many athletes. During the video, there will be playback of wonderful moments, slow motion

commentary, judges' ranking, and other links. In the massive videos, athletes are taken as the unit to cut, and only the complete sets of aerobics movements are reserved. 298 videos were included in the training data, and 45 videos were included in the test data.

4.3. Evaluation Index

(1) Accuracy is the most common performance metric in classification models. It is suitable for two-class models and can also be used for multiclass models. The calculation of accuracy is also relatively simple. Assuming that the classification model is g and the test set contains N data in D, the accuracy rate calculation formula is as follows:

$$A = \frac{1}{N}\sum_{i=1}^{N}\left(f\left(a_i\right) = \text{label}_i\right). \tag{11}$$

(2) Bleu is currently the closest indicator to a human score. Bleu adopts the matching principle of N-gram. N-gram represents a sentence as a sequence of n consecutive words. When $N = 1$, the result is Bleu 1.

4.4. Experimental Results. The ultimate goal of this article's multiclassification is to realize the automatic recognition of videos. This article takes the average value of Blue 1 to Blue 4, Blue, as the evaluation index and compares the recognized description of aerobics with the correct description. The experimental results are shown in Table 1. It can be clearly seen that the method used in this article to automatically recognize aerobics videos using the conversion of video multilabel classification has outstanding performance. In addition, Figures 6–8 also show the visual results of the experiment.

4.5. Ablation Study. The experimental results as shown in Table 2 compared from the table are the average values of Bleu 1 to Bleu 4, Blue. You can see three data sets in the table. Two of them are self-built. The two data sets are different when labeling the video description. Ours A is the most direct natural language due to the decomposition of aerobics. The professional requirements of description are very high. Ours B adjusts the description sentence according to professional terminology when describing the mark. From the experimental results of these three data sets, it can be seen that the model with the attention mechanism introduced in this article has better performance regardless of the experimental results on the MSVD data set or the experimental results on the self-built data set. The use of planned sampling in the experiment can also improve the experimental results to a certain extent.

Table 1: Classification accuracy of each category.

Category	Acc
Result 1	0.75
Result 2	0.75
Result 3	0.68
Result 4	0.72
Result 5	0.69
Result 6	0.68
Result 7	0.73
Result 8	0.73
Result 9	0.78
Result 10	0.80

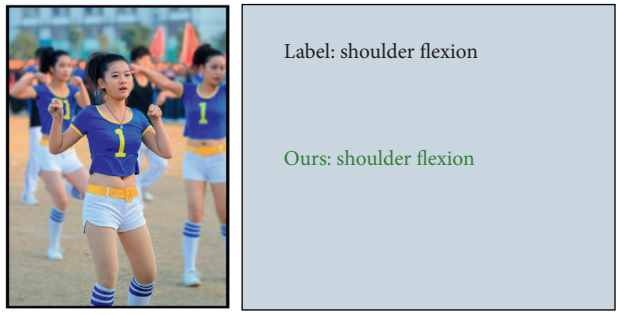

Figure 6: Visualization results of test example 1.

Figure 7: Visualization results of test example 2.

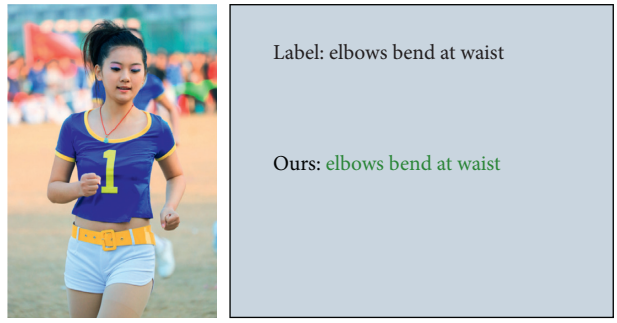

Figure 8: Visualization results of test example 3.

Table 2: Results of ablation research experiments.

Method	S2VT	Attention	Plan to sample
MSVD	17.2	17.9	18.8
Ours A	8.7	10.2	11.6
Ours B	10.9	11.3	13.2

5. Conclusion

In this paper, we will combine computer vision and deep learning related knowledge to realize the intelligent recognition and representation of specific human movements in aerobics video sequences. Therefore, this article proposes an automatic recognition method for floor exercise videos based on three-dimensional convolutional networks and multilabel classification. Since two-dimensional CNN loses time information when extracting features, this paper uses three-dimensional convolutional networks to perform video recognition. The feature is taken in time and space, and the extracted features are subjected to multiple binary classifications to achieve the goal of multilabel classification. We will conduct comparison and simulation experiments, and the experimental results prove the effectiveness and superiority of our algorithm.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. A. R. Ahad, *Computer Vision and Action Recognition: A Guide for Image Processing and Computer Vision Community for Action Understanding*, Springer Science & Business Media, Berlin, Germany, 2011.

[2] S. Arseneau and J. R. Cooperstock, "Real-time image segmentation for action recognition," in *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pp. 86–89, Victoria, Canada, Augest 1999.

[3] I. T. Toudjeu and J. R. Tapamo, "Slope pattern spectra for human action recognition," in *Proceeding of the International Conference Image Analysis and Recognition*, pp. 381–389, Springer, Póvoa de Varzim, Portugal, June 2018.

[4] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, In press.

[5] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, 2020, In press.

[6] X. Ning, K. Gong, W. Li, and L. Zhang, "JWSAA: joint weak saliency and attention aware for person re-identification," *Neurocomputing*, vol. 453, pp. 801–811, 2021.

[7] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4768–4777, Honolulu, HI, USA, July 2017.

[8] S. Pal, P. K. D. Pramanik, T. Majumdar, and P. Choudhury, "A semi-automatic metadata extraction model and method for video-based e-learning contents," *Education and Information Technologies*, vol. 24, no. 6, pp. 3243–3268, 2019.

[9] J. Dong, X. Li, C. Xu et al., "Dual encoding for zero-example video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9346–9355, Long Beach, CA, USA, June 2019.

[10] Y. Zhang, Y. Chen, X. Bai et al., "Adaptive unimodal cost volume filtering for deep stereo matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12926–12934, New York, NY, USA, February 2020.

[11] L. Zhou, X. Bai, X. Liu, J. Zhou, and E. R. Hancock, "Learning binary code for fast nearest subspace search," *Pattern Recognition*, vol. 98, Article ID 107040, 2020.

[12] C. Wang, X. Wang, X. Bai, Y. Liu, and J. Zhou, "Self-supervised deep homography estimation with invertibility constraints," *Pattern Recognition Letters*, vol. 128, pp. 355–360, 2019.

[13] M. Woitas, ""Exercise teaches you the pleasure of discipline"–the female body in jane fonda's aerobics videos," *Historical Social Research/Historische Sozialforschung*, vol. 43, no. 2, pp. 148–164, 2018.

[14] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Computer Vision*, vol. 12, no. 1, pp. 3–15, 2017.

[15] D. Alfermann, M. J. Lee, and S. Würth, "Perceived leadership behavior and motivational climate as antecedents of adolescent athletes' skill development," *Athletic Insight: Online Journal of Sport Psychology*, vol. 7, no. 2, pp. 14–36, 2005.

[16] Z. Chu, M. Hu, and X. Chen, "Robotic grasp detection using a novel two-stage approach," *ASP Transactions on Internet of Things*, vol. 1, no. 1, pp. 19–29, 2021, In press.

[17] C. Yan, G. Pang, X. Bai et al., "Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss," *IEEE Transactions on Multimedia*, 2020.

[18] Y. Ding, X. Zhao, Z. Zhang, W. Cai, and N. Yang, "Multiscale graph sample and aggregate network with context-aware learning for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4561–4572, 2021.

[19] J. Zhang, Y. Liu, H. Liu, and J. Wang, "Learning local–global multiple correlation filters for robust visual tracking with Kalman filter redetection," *Sensors*, vol. 21, no. 4, p. 1129, 2021.

[20] X. Ning, X. Wang, and S. Xu, "A review of research on co-training. concurrency and computation: practice and experience," *Journal of Healthcare Engineering*, 2021, In press.

[21] L. R. Medsker and L. C. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, 2001.

[22] X. Guo, H. Zhang, L. Ye, and S. Li, "An approach to learning users' intention to legal consultation with normalized tensor decomposition and BI-LSTM," *Computers, Materials & Continua*, vol. 63, no. 1, pp. 315–336, 2020.

[23] M. Hasnain, S. R. Jeong, M. F. Pasha, and I. Ghani, "Performance anomaly detection in web services: an rnn-based approach using dynamic quality of service features," *Computers, Materials & Continua*, vol. 64, no. 2, pp. 729–752, 2020.

[24] G. Yang, J. Zeng, M. Yang, Y. Wei, and X. Wang, "Ott messages modeling and classification based on recurrent neural networks," *Computers, Materials & Continua*, vol. 63, no. 2, pp. 769–785, 2020.

[25] Y. Tong, L. Yu, S. Li, J. Liu, H. Qin, and W. Li, "Polynomial fitting algorithm based on neural network," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 1, no. 1, pp. 32–39, 2021.

[26] X. Ning, Y. Wang, W. Tian, L. Liu, and W. Cai, "A biomimetic covering learning method based on principle of homology continuity," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 1, no. 1, pp. 9–16, 2021.

[27] Z. Huang, P. Zhang, R. Liu, and D. Li, "Immature apple detection method based on improved Yolov3," *ASP Transactions on Internet of Things*, vol. 1, no. 1, pp. 9–13, 2021.

[28] D. Zhu, Y. Sun, X. Li, and R. Qu, "Massive files prefetching model based on LSTM neural network with cache transaction strategy," *Computers, Materials & Continua*, vol. 63, no. 2, pp. 979–993, 2020.

[29] W. Fang, F. Zhang, Y. Ding, and J. Sheng, "A new sequential image prediction method based on lstm and dcgan," *Computers, Materials & Continua*, vol. 64, no. 1, pp. 217–231, 2020.

[30] B. Yan, X. Tang, J. Wang, Y. Zhou, and G. Zheng, "An improved method for the fitting and prediction of the number of Covid-19 confirmed cases based on LSTM," *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1473–1490, 2020.

[31] Z. Qu, B. Cao, X. Wang et al., "Feedback lstm network based on attention for image description generator," *Computers, Materials & Continua*, vol. 59, no. 2, pp. 575–589, 2019.

[32] W. Cai, B. Liu, Z. Wei, M. Li, and J. Kan, "TARDB-net: triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 11291–11312, 2021.

[33] R. Liu, X. Ning, W. Cai, and G. Li, "Multiscale dense cross-attention mechanism with covariance pooling for hyperspectral image scene classification," *Mobile Information Systems*, vol. 2021, Article ID 9962057, 15 pages, 2021.